# Domain Question

* Imagine you are a statistician working at PG&E and want to accurately predict when and where a fire will start
* What is the end goal? Actionable intel?
* Refinements to the question
* What is the response

# Data Collection

* Experimental design
* What features/variables/measurements to collect and when
* Sampling bias
* Existing data
* How would you organize data
* How to measure the response
* Are samples iid given X
* Feature engineering

# Data Cleaning

* Consistent units/resolutions
* Missing values (are they missing at random)
* Nonsensical observations/inconsistent/erroneous
* Duplicates
* Time synchronization
* Filtering observations/variables
  + Remove variables with 0 variance
* Categorical variables
* Transforming data (log(x+1)) or aggregating/normalization
* Centering/scaling
* **Data splitting**

# EDA

* Tools: scatterplots, boxplots, kernel densities, heatmaps, pair plots, histograms
* Dimensionality reduction: PCA, NMF, ICA
* Relationships between features (pair plots)
* Overplotting: transparency, point sizes, subsampling

# Modeling

* Method model
* Can always apply the method, but if you want to do statistical inference, need to carefully check model assumptions (though understanding the implicit assumptions when applying the method allows for critical understanding of the results)
* Correlated features
* Bias-variance tradeoff
* Unsupervised vs supervised learning

# Post-Hoc Analysis

* Look at residuals/errors or patterns among them
* Heatmaps
* Diagnostic plots

# Interpretation of Results

* Feature importances/stability
* Connecting with the domain

# Update Domain Knowledge